3С

УДК 004.934

М.М. Сажок[1,2], В.В. Робейко[1]
[1,2]Міжнародний науково-навчальний центр інформаційних технологій та систем
просп. академіка Глушкова, 40, м. Київ, 03680, Україна
[2]Інститут кібернетики ім. В.М. Глушкова, м. Київ, Україна
просп. Академіка Глушкова, 40, м. Київ, 03680, Україна

Морфологічний розклад слів на основі лексичного наголосу в задачах розпізнавання українського мовлення

M. Sazhok[1,2], V.V. Robeiko[1]
[1,2]International Research and Training Center of Information Technologies and Systems
40 prospekt akademika Hlushkova, Kyiv, 03680, Ukraine
[2]Hlushkov Institute of Cybernetics
40 prospekt Akademika Hlushkova, Kyiv, 03680, Ukraine

Lexical Stress-based Morphological Decomposition for Ukrainian Speech Recognition Tasks

Н.Н. Сажок[1,2], В.В. Робейко[1]
[1,2]Международный научно-учебный центр информационных технологий и систем
просп. академика Глушкова, 40, г. Киев, 03680, Украина
[2] Институт кибернетики им. В.М. Глушкова, г. Киев, Украина
просп. Академика Глушкова, 40, г. Киев, 03680, Украина

Морфологическое разложение слов на основании лексического ударения в задачах распознавания украинской речи

У статті описано новий метод морфологічного розкладу слів шляхом моделювання лексичного наголосу, що актуально для систем розпізнавання українського мовлення. Критерій сегментації формулюється на підставі великого текстового корпусу та слів із позначеним наголосом. Наведений алгоритм пошуку знаходить одну або декілька найбільш імовірних сегментацій. Описуються експериментальні дослідження, обговорюються результати та плани на майбутнє.
Ключові слова: лексичний наголос, морфологічний розклад, розпізнавання українського мовлення.

This paper presents an approach to the morphological level word segmentation based on lexical stress modeling, which is prospective for Ukrainian speech recognition systems. The formulated segmentation criterion is based on a training set of words with manually pointed stresses and a large text corpus. The described search algorithm finds one or more segmentations with the best likelihood. The developed toolkit is presented, experimental research is described and results are discussed.
Key words: lexical stress, morphological decomposition, Ukrainian speech recognition.

В статье описан новый подход к морфологическому разложению слов на основе моделирования лексического ударения, что актуально для систем распознавания украинской речи. Критерий сегментации формулируется на основании большого текстового корпуса и слов с обозначенным ударением. Приведенный алгоритм поиска находит один или несколько наиболее вероятных сегментаций. Описываются экспериментальные исследования, обсуждаются результаты.
Ключевые слова: лексическое ударение, морфологическое разложение, распознавание украинской речи.

Вступ

Явище лексичного наголосу відіграє важливу роль у багатьох мовах. Наголошені та ненаголошені фонеми в українській мові відрізняються за багатьма просодичними параметрами. Тому під час генерування мовленнєвого сигналу за текстом необхідно прогнозувати лексичний наголос у словах. Наголос для відомих слів береться зі словника. Частка слів, які не входять до словника, тобто OOV-слів (від англ. *out of vocabulary*), може складати суттєвий відсоток у текстах за рахунок рідковживаних слів, термінології, власних назв, слів із помилками тощо. Наголошені фонеми майже завжди вимовляються відповідно до правил вимови, навіть у спонтанному мовленні. І цю властивість можна використати в задачах розпізнавання.

Проблемі прогнозування наголосу присвячено багато наукових досліджень. У [1] автори припускають, що морфологічний розклад для прогнозування лексичного наголосу особливо корисний у випадках недостатності локального контексту. Представлення слів як послідовності певним чином обґрунтованих сегментів або морфем є ключем до моделювання словотвору та до виходу за межі словникової моделі лексикону. Відомі методи морфологічного розкладу покладаються виключно на орфографію [2], [3]. У наших дослідженнях прогнозування лексичного наголосу та морфологічний розклад розглядаються як результат одного і того ж процесу, через який на основі орфографічного написання виявляються фонетичні, синтаксичні та семантичні ознаки.

В українській мові позиція наголосу є нерегулярною та може змінюватися навіть у формах одного і того ж слова та в однокореневих словах (наприклад: *фо́то – фото́граф – фотогра́фія – фотографу́є – фотографува́ти*). Завдяки доступу до лексикографічної системи [4], ми отримали можливість аналізувати понад 1,8 млн описаних експертами словоформ із позначеним лексичним наголосом. Створений без посередньо авторами базовий текстовий корпус містить 275 млн неперевірених реалізацій слів, що складають словник із близько двох мільйонів словоформ. Половина слів словника цього корпусу описана в лексикографічній системі. Частка корпусу, не відображена в лексикографічній системі, складає 2,5%, які ми фіксуємо як початковий показник OOV. Додавання 200 тисяч найбільш частотних слів до словника дало змогу скоротити показник OOV до 0,5%. Таким чином, прогнозування наголосів сприятиме позиціюванню лексичного наголосу для величезної кількості як нових, так і відомих системі слів.

Причина введення наголосів у системах озвучення текстів є очевидною через необхідність генерувати звуковий сигнал, що відповідає людському сприйняттю таких просодичних ознак, як тривалість, висота основного тону та енергія сигналу. У задачах розпізнавання мовлення моделі переходу в простір первинних ознак загалом є інваріантними до просодичних ознак. Утім, ми вважаємо, що введення як наголошених, так і ненаголошених фонем до алфавіту української мови є суттєвим з огляду на фонетичні, лексичні й акустичні факти. Наголошені голосні у багатьох випадках діють як окремі фонеми, змінюючи граматичну функцію слова та його значення у більш ніж 5% слів, що спостерігаються в базовому текстовому корпусі (явище омографії).

Методи перетворення графем на фонеми, подібні до [5], також можуть напряму застосовуватися для моделювання лексичного наголосу, хоча описаний у згаданій роботі підхід не передбачає врахування структурних властивостей наголосу.У цьому дослідженні ми пропонуємо зосередитись на моделюванні властивостей наголосу, а вже потім перетворювати текст із наголосами на послідовності фонем методами, описаними, наприклад, у [6], які дають змогу враховувати особливості вимовляння. У реалізаціях згаданого методу достатньо задати 30 правил типу *знайти-замінити-та-змінити-позицію* для перетворення графемного тексту на фонемний, що моделює базову українську вимову.

Використання інформації про наголос у задачах розпізнавання мовлення

Щоб дослідити акустичний аспект лексичного наголосу, ми оцінили параметри акустичної моделі, розглядаючи наголошені та ненаголошені голосні як різні фонеми та проаналізували відмінності між ними за допомогою інструментарію візуалізації прихованих марківських моделей [7]. На рис. 1 показана відмінність між акустичними моделями ненаголошених та наголошених фонем **а** та **і**, параметри яких оцінені на 40-годинному відрізку акустичного корпусу українського мовлення [8].

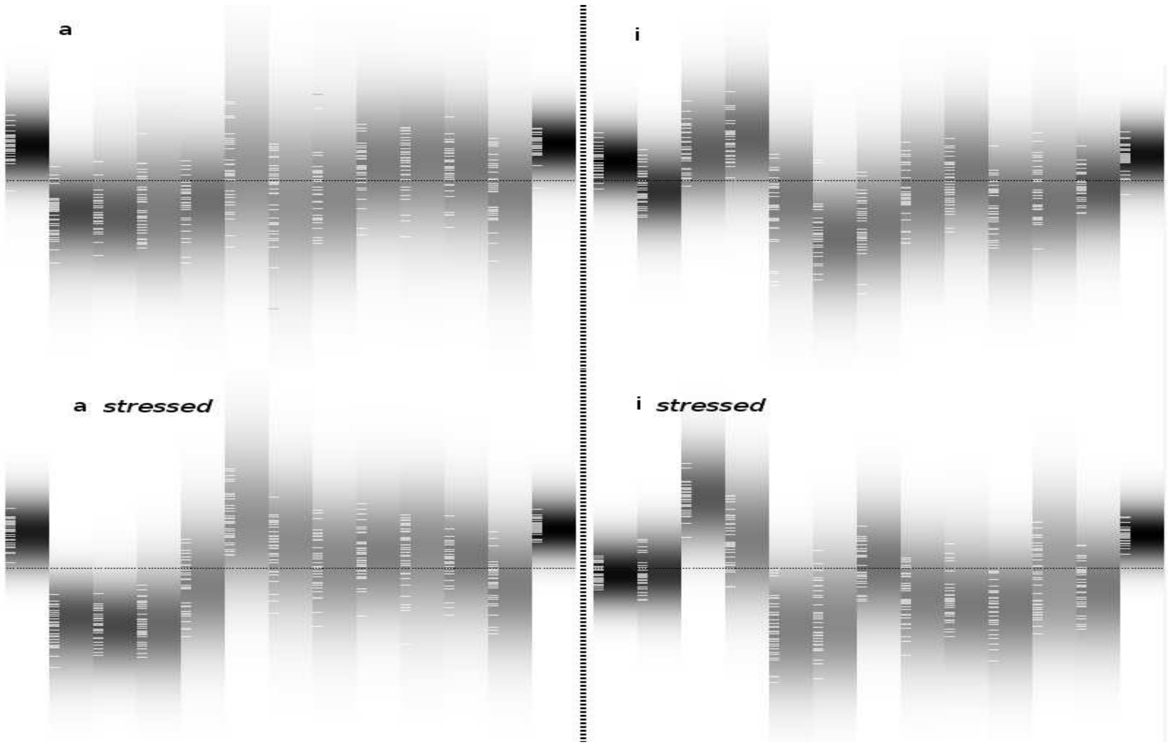


Рисунок 1 – Візуалізація акустичних моделей фонем **а** та **і**

На рисунку представлені області, у яких спостерігаються центральні стани фонем. Ці області апроксимуються сумішшю із 32 нормальних законів у просторі первинних ознак на основі мел-кепстральних коефіцієнтів із застосуванням віднімання середнього, що загалом складає 13-вимірний вектор. Лінія з крапок відповідає нульовому значенню. Візуально наголошені моделі виглядають як підмножини для більшості вимірів. Накладання областей, а не їх включення, найбільш яскраво спостерігається у п'ятому вимірі моделі фонеми **а** і в першому вимірі фонеми **і**. На веб-сторінці інструментарію [7] можна ознайомитися з іншими акустичними моделями.

Аналізуючи матриці переходів між станами, ми бачимо, що значення, які відповідають робочим (емітентним) станам, у півтора – два рази більші для моделей наголошених фонем. Цей факт підтверджує суттєву відмінність довжин фонем залежно від наголошеності.

Введення як наголошених, так і ненаголошених голосних фонем для розпізнавання української мови є відносно невеликою додатковою витратою обчислювальних ресурсів

(у результаті маємо 6 ненаголошених голосних та 6 наголошених – **а**, **е**, **у**, **і**, **о**, **и**, **А**, **Е**, **У**, **І**, **О**, **И**). Проте подібне розширення алфавіту фонем є суттєвим для мов, що мають значно більшу кількість голосних фонем.

Найбільш переконливі аргументи за або проти введення наголошених фонем надає аналіз попередніх результатів розпізнавання українського мовлення.

Для оцінки параметрів акустичної моделі використовувалися багатодикторна й однодикторна навчальні вибірки для обох версій алфавіту фонем на 49 та 55 монофонів відповідно. На лексичному рівні при послівному розпізнаванні злитого мовлення використовувалися бі- та триграмні моделі, а для пофонемного та поскладового розпізнавання допускався вільний порядок слідування елементів. Для того, щоб можна було порівнювати результати, ми ігнорували інформацію про наголос у розпізнаних послідовностях слів і фонем. У всіх випадках спостерігалися результати, кращі на 12 – 23% для акустичних моделей з ненаголошеними та наголошеними голосними щодо послівної або пофонемної помилки.

Слід зазначити, що перевагою морфологічного розкладу є можливість представити весь лексикон системи розпізнавання за допомогою практично незмінної множини сегментів морфемного рівня.

Модель сегментації слів на основі лексичного наголосу

Нехай маємо словник W, що містить слова з позначеними атрибутами, такими як лексичний наголос. Кожне слово w зі словника W може бути розкладене на послідовність символів $q^{(w)} = \left(q_1, q_2, \ldots, q_k, \ldots, q_{K_w}\right)$, які містяться в алфавіті літер або фонем Q.

Ми розглядаємо послідовності $q^{(w)}$ як сегменти деякої сегментації $s^{(w)}$ серед усіх допустимих сегментацій $S^{(w)}$ слова w, причому i-й сегмент сегментації $s^{(w)}$

$$s_i^{(w)} = \left(s_{i1}, s_{i2}, \ldots, s_{ij}, \ldots, s_{iL_i^{(w)}}\right) \quad (1)$$

разом із іншими сегментами $s^{(w)}$ покривають усю $q^{(w)}$ без перекриттів, що означає, що для будь-якої $w \in W$

$$\sum_i L_i^{(w)} = K_w,\ 1 \le L_i^{(w)} \le \min\{L_{\max}, K_w\}, \quad (2)$$

$$I\left(s_{11}^{(w)}\right) = 1 \text{ та } I\left(s_{i1}^{(w)}\right) = I\left(s_{(i-1)L_{i-1}^{(w)}}^{(w)}\right) + 1,\ i > 1, \quad (3)$$

де $I(\cdot)$ повертає індекс елемента сегменту в $q^{(w)}$. Обмеження на найбільшу довжину сегмента, $L_{\max}$, визначає порядок моделі. Також можуть бути введені й інші обмеження на сегментування, наприклад, заборона на два поспіль склади, наголошені основним наголосом.

Об’єднуючи всі сегменти допустимих сегментацій для всіх слів зі словника W, ми формуємо множину сегментів

$$S = \bigcup_{w \in W,\ s^{(w)},\ i} s_i^{(w)} \tag{4}$$

і розглядаємо кожен сегмент s_i у цій множині, не зважаючи на належність до слів.

Рівень наголосу $\theta_k^{(w)} = \{0,1,2\}$, який приписується кожному символу, формує відповідну послідовність атрибутів $\theta^{(w)} = (\theta_1, \theta_2, ..., \theta_k, \theta_{K_W})$. Ми припускаємо, що відмінний від нуля рівень наголосу може відповідати символам, якими вводиться склад, принаймні потенційно. Зазвичай, такими символами є голосні, доповнені специфічними приголосними, такими як «**r**» у словенській мові [3]. Для інших символів рівень наголошеності не допускається, а тому завжди дорівнює нулеві. Значення рівнів наголосу можуть бути обмежені нулем або одиницею, що означає, що розглядається лише основний наголос. Допускається введення інших значень, що відповідають різним атрибутам символів, які можуть бути прихованими на письмі (риски, крапки, коронки тощо), та комбінаціям цих атрибутів. Отже, в загальному випадку ми посилаємося на $\theta^{(w)}$ як на послідовність атрибутів для відповідних символів у слові w.

Очевидно, індекс, який повертається у (3) є одним і тим же, що і для $\theta^{(w)}$, чиї підпослідовності відповідають $s_i^{(w)}$. Послідовності атрибутів, що відповідають сегментації $s^{(w)}$, у свою чергу, формують множину $\Theta^{(w)}$.

Ми можемо оцінити ймовірність послідовності атрибутів θ за умови сегмента s_i, який спостерігався в навчальній вибірці:

$$P(\theta \mid s_i) \approx \frac{c(s_i, \theta)}{c(s_i)}, \tag{5}$$

де $c(s_i, \theta)$ є кількістю сегментів s_i з атрибутом наголосу, що визначений індикатором наголосу θ, а $c(s_i)$ – загальна кількість s_i. Усі підрахунки здійснюються за текстовим корпусом для слів, що входять до словника наголосів. Для сегментів з малою частотою доцільно застосувати методику згладжування.

Остаточно здійснюється пошук за всіма допустимими сегментаціями $s^{(w)}$ та послідовностями атрибутів θ, що задовольняють вираз:

$$\left(\hat{s}^{(w)}, \hat{\Theta}^{(w)}\right) = \underset{s^{(w)}, \Theta^{(w)}}{\operatorname{argmax}} \prod_{i,\theta} P\left(\theta \mid s_i^{(w)}\right). \tag{6}$$

У словах, які належать словнику наголосів, θ визначається для кожного сегменту $s_i^{(w)}$ однозначно, в іншому випадку пошук здійснюється засобами динамічного програмування для всіх допустимих послідовностей атрибутів.

Таким чином, щоб виконати морфологічний розклад, ми ввели модель сегментації за ознаками, що, як правило, не відображаються в орфографії. До цих ознак відноситься лексичний наголос. Не кожний отриманий сегмент може бути допустимою морфемою внаслідок потенційно більш строгих обмежень на вміст морфеми, таких як наявність принаймні однієї голосної фонеми. Ці обмеження можна обійти шляхом об'єднання сегментів із одним або кількома прилеглими сегментами.

Аналіз графу сегментації

Ми сконструювали граф динамічного програмування, на якому знаходження найкоротшого шляху еквівалентно пошуку (6). Кожний вхідний символ вводить множину допустимих пар (*сегмент, атрибути*), що розташовані у вузлах графа і де накопичується частковий критерій. Запам'ятовуючи *N* перспективних стрілок, що входять у вузли, ми можемо отримати *N* кращих сегментацій слова.

На рис. 2 показано приклад пошуку найкращого прогнозу наголосів (6) для власної назви *Обама*, що відсутнє у базовому словнику наголосів. Слово представлене як конкатенація всіх допустимих сегментів символів, де довжина найдовшого сегмента обмежується п'ятьма символами. Вхідні символи переведені у нижній регістр, додано символ «|», що позначає межі слів. Допустимі сегменти з атрибутами, які вводяться поточним спостережуваним елементом, будемо подавати в компактній формі, одразу відображаючи результат дії атрибутів. Так запис «**обАм**» у п'ятій колонці, який назвемо *іменем вузла*, означає сегмент (*о*, *б*, *а*, *м*) під дією вектора атрибутів (0, 0, 1, 0). Потенційно оптимальні стрілки або показуються або кодуються іменем попереднього вузла. Позначені часткові критерії ґрунтуються на логарифмі ймовірності. Оптимальна траєкторія, відповідні вузли та критерії виділені потовщенням.

На цьому прикладі ми ілюструємо заборону на слідування двох поспіль наголошених сегментів: у 7-й колонці сегмент «**мА**» слідує за сегментом «**а**», а не «**обА**». Оскільки не вводиться обмежень на вміст сегментів, допускається сегмент, що містить одну приголосну «**б**», як у третій колонці. Таким чином ми гарантуємо успішність пошуку (6) для будь-якого слова. Система може вирішити, що обидва прилеглі сегменти належать до єдиної морфеми залежно від обмежень, які накладає експерт. Щоб сформувати формально допустиму морфему, ми можемо приєднати сегмент «**б**» до попереднього сегмента, віддаючи перевагу більш частотній морфемі та приходячи до сегментації *Об-а́ма*. Можемо побачити, що це слово іноземного походження апроксимується морфемами з рідної мови. Модель, поновлена зразками автоматично наголошених нових слів, отримує змогу навчитись на нові морфеми, що потенційно може привести до лінгвістично більш обґрунтованого розкладу даного слова та його форм у вигляді: *Оба́м-а*, *Оба́м-и* тощо.

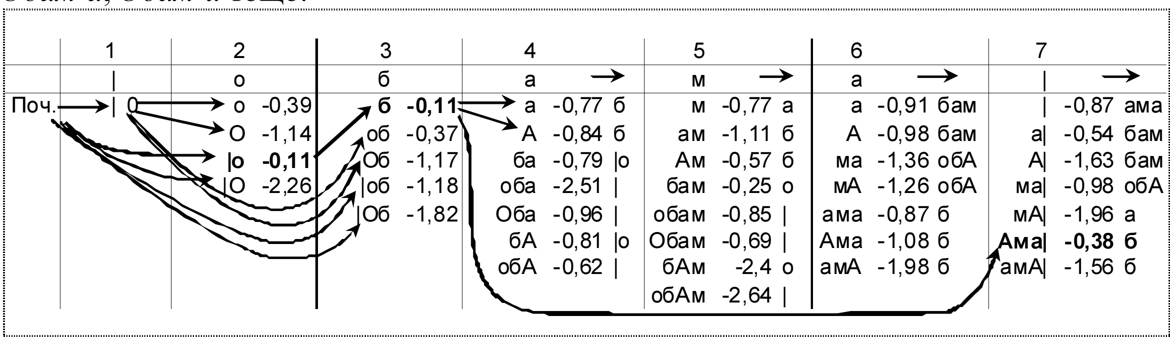


Рисунок 2 – Граф прогнозування наголосу для слова *Обама*, яке відсутнє в базовому словнику наголосів

Реалізація системи прогнозування наголосів

Для реалізації описаного алгоритму сегментації слів було розроблено інструментарій, що складається із трьох модулів. На даний момент допускається оперування лише інформацією про основний лексичний наголос.

Перший модуль – *putstress* – готує дані, необхідні для обчислення ймовірностей (5) за вхідною базою даних і знань, частотним словником та, при потребі, за скоригованими пропорціями частот омографів. Модуль намагається отримати знання щодо позиції наголосу для кожного слова та, в разі успіху, зберігає слова, доповнені позначкою наголосу та частотою в окремий файл. Знайдені омографи зберігаються з частотами, уточненими відповідно до їх скоригованих пропорцій, які експерт може знову ж таки коригувати з наступним повторним запуском цього модуля.

Другий модуль – *guessstress* – реалізує процедуру пошуку (6), отримуючи N кращих послідовностей сегментів із відповідними атрибутами. Частотний словник слів із позначеними наголосами є вхідними даними для оцінки ймовірностей гіпотетичних підпослідовностей символів.

Третій модуль – *prep_stressvcb* – формує словник наголосів за отриманими попереднім модулем сегментаціями. Декілька допоміжних модулів дають змогу виокремити різноманітну інформацію із вхідних даних, оцінених моделей та сегментацій. Усі модулі написані мовою *Perl*.

Опис текстових даних

Словник наголосів отримано з підмножини електронної лексикографічної системи, що містить 151 962 лем, включаючи понад десять тисяч імен, що загалом становить 1,90 млн словоформ [4]. Внаслідок аналізу спільної орфографії, кількість слів, що мають або відмінне написання, або основний наголос, складає 1,83 млн.

Базовий текстовий корпус отримано з гіпертекстових даних, завантажених із ряду веб-сайтів, що містять новини та публіцистику (60%), художню літературу (8%), енциклопедичні статті (24%) та юридичний матеріал (8%). Зазначимо, що дані, завантажені з новинних сайтів, містять численні коментарі користувачів, які ми розглядаємо як текстові реалізації спонтанного мовлення. Надалі ми посилатимемось на базовий текстовий корпус, як на корпус 275M. Відповідно до наведеної характеристики цього корпусу в табл. 1, ми спостерігаємо в середньому 6,64 словоформ на лему, тоді як цей показник удвічі більший для словника на основі [4] і становить 12,3. Додавши до відомих слів словника найбільш частотних 200 тис. слів, ми скоротили показник OOV до менше ніж 0,5%.

Таблиця 1 – Характеристика базового корпусу 275M

Кількість слів	Кількість речень	Словник			OOV	Кількість омографів
		Усі слова	Відомі слова	Відомі леми		
275 288 408	1 752 371	1 996 897	801 040	120 554	2,51%	16 729 476

Ми бачимо, що частка слів-омографів, які мають дві та більше допустимих позицій наголосу, складає 6% від тексту. Зауважимо, що омографи можуть мати різну частоту, що впливає на частоту певних сегментів. Тому експерту надано можливість коригувати пропорції частоти омографів, словник яких складається з понад 14 000 елементів.

Експериментальні дослідження

Відомі слова та OOV-слова були досліджені окремо. Метою дослідження відомих слів було з’ясувати, наскільки значна частина словника може бути закодована без за-

значення інформації про лексичний наголос. Найбільший порядок L_{max} моделі рівний п'яти, багатозначність було обмежено чотирма кращими сегментаціями, за якими формувався словник наголосів. Експерт скоригував пропорції частотності для перших за частотою 500 омографів.

Системою виявлено близько мільйона пар (*сегмент, наголос*). Частоти для сегментів різної довжини показані у табл. 2.

Таблиця 2 – Кількісні характеристики виявлених сегментів

Довжина сегмента, *L*	1	2	3	4	5
Кількість сегментів	46	1 781	35 280	233 816	721 575
Частота (млн)	2 115,652	1 848,766	1 581,879	1 314,993	1 070,579

Було використано 215 000 сегментів для передбачення наголосу у словах кор.пусу 275М. Для менше ніж 1% відомих слів наголос було передбачено хибно. Визначення наголосу для 5 000 OOV-слів дало помилку у 21,1% слів, що відповідає 5,3% складів. Варто зазначити, що більше половини неправильно визначених наголосів припадає на рідкісні запозичення з інших мов.

Чи не найбільший інтерес викликає реакція системи на рух наголосу в однокореневих словах. Перевіривши слова, похідні від *фото*/*фотографія*, ми виявили, що лише слово *фотограф* мало хибно визначений наголос.

Висновки

Запропонована модель сегментації морфемного рівня дає змогу одночасно виявляти ознаки, які, як правило, ігноруються при написанні слів. Введена багатозначність дає змогу обирати кращу гіпотезу з урахуванням ширшого контексту на рівні слів, що є актуальним при аналізі омографів. Подальше вдосконалення запропонованої моделі полягає у введенні контексту на сегментному рівні.

Оцінювання параметрів моделі передбачає покращення сили прогнозування за рахунок додання до навчальної вибірки невідомих слів та коригування експертом наголосів у словах між ітераціями. Необхідно передбачити інтерактивну процедуру такого коригування, щоб уникати зайвої роботи з однокореневими словами під час аналізу. Планується також дослідити вплив вибору порядку моделі, ввести нові ознаки, використати фонемний вхідний текст та розширити коло досліджуваних мов. Зважаючи на доступність реалізації підходу [5] у відкритому коді, існує можливість провести порівняльний аналіз обох методів на одному й тому ж матеріалі.

Література

1. Black A. Issues in Building General Letter to Sound Rules / A. Black, K. Lenzo, V. Pagel // 3rd ESCA Workshop on Speech Synthesis. – Australia : Jenolan Caves, 1998. – P. 77-80.
2. Creutz Mathias. 2004. Induction of a simple morphology for highly-inflecting languages / Creutz Mathias, Lagus, Krista // In : Proc. 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON). – Barcelona. - P. 43–51.
3. Automatic lexical stress assignment of unknown words for highly inflected Slovenian language / [Gams Matjaž et al.] // In : Text, Speech and Dialogue. Springer Berlin Heidelberg. – 2006. – P. 165-172.
4. [Электронный ресурс]. – Ресурс доступа : http://lcorp.ulif.org.ua/dictua/
5. Bisani M. Joint-Sequence Models for Grapheme-to-Phoneme Conversion / M. Bisani, H. Ney // Speech Communication. – May 2008. – Vol. 50, Issue 5, - P. 434-451.
6. Robeiko V. Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian / V. Robeiko, M. Sazhok // In Proc. UkrObraz'2012. – Kyiv, 2012. – P. 43-46.
7. [Электронный ресурс]. – Ресурс доступа : www.cybermova.com/speech/visual-hmm.htm
8. Ukrainian Broadcast Speech Corpus Development / [Valeriy Pylypenko, Valentyna Robeiko, Mykola Sazhok, et al.] // In Proc. Specom'2011. – Kazan : RF. – P. 244-247.

Literatura

1. Black A. Issues in Building General Letter to Sound Rules . 3rd ESCA Workshop / Black A., Lenzo K., Pagel V.
2. Creutz Mathias. Induction of a simple morphology for highly-inflecting languages / Creutz Mathias, Lagus Krista // In: Proc. 7th Meeting of the ACL SIGPHON. – 2004.
3. Automatic lexical stress assignment of unknown words for highly inflected Slovenian language / [Gams Matjaž et al.] // In: Text, Speech and Dialogue. Springer Berlin Heidelberg, 2006.
4. http://lcorp.ulif.org.ua/dictua/
5. Bisani M. Joint-Sequence Models for Grapheme-to-Phoneme Conversion / M. Bisani, H. Ney // Speech Communication.
6. Robeiko V. Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian // V. Robeiko, M. Sazhok // In Proc. UkrObraz'2012.
7. www.cybermova.com/speech/visual-hmm.htm
8. Ukrainian Broadcast Speech Corpus Development / [Valeriy Pylypenko, Valentyna Robeiko, Mykola Sazhok, et al.] // In Proc. Specom'2011.

RESUME

M. Sazhok, V.V. Robeiko

Lexical Stress-based Morphological Decomposition for Ukrainian Speech Recognition Tasks

This paper presents an approach to word morphological decomposition based on lexical stress modeling. Lexical stress prediction and morphological decomposition are considered as a result of the same process through which phonetic, syntactic and semantic hidden features can be discovered from word spelling.

Given motivation confirms that introduction of both stressed and unstressed vowels to the speech recognition system phoneme alphabet, at least for Ukrainian, is essential due to phonetic, lexical, and acoustical facts.

Word segmentation quality is estimated by a hidden variable that assigns the lexical stress. The formulated segmentation criterion is based on a training set of words with manually pointed stresses and a large text corpus. The described search algorithm finds one or more segmentations with the best likelihood by means of dynamic programming.

The developed toolkit allows for assigning a primary lexical stress in unknown words. Beside required input text data and basic stress vocabulary, an expert may provide homograph occurrence proportions, which is essential for operating with correct word segment frequency. The experimental research is described as well as results and future plans are discussed.

Стаття надійшла до редакції 10.06.2013.